Bernd Malle

The Name of the Event (Workshop/Symposium)

06.03.2016

Interactive Machine Learning for improving K-anonymity

b.malle@hci-kdd.org

Overview / Contents

1. What is Machine Learning?
2. What is interactive Machine Learning?
3. What is k-anonymity?
    - Privacy in the 21st century…?
4. Influence on k-anonymity on ML performance
5. Can we improve this via iML?
6. What is Graphinius?
7. Structure of experiments in AK-HCI

What is Machine Learning

- Definition by Tom Mitchell:

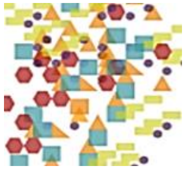
> "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

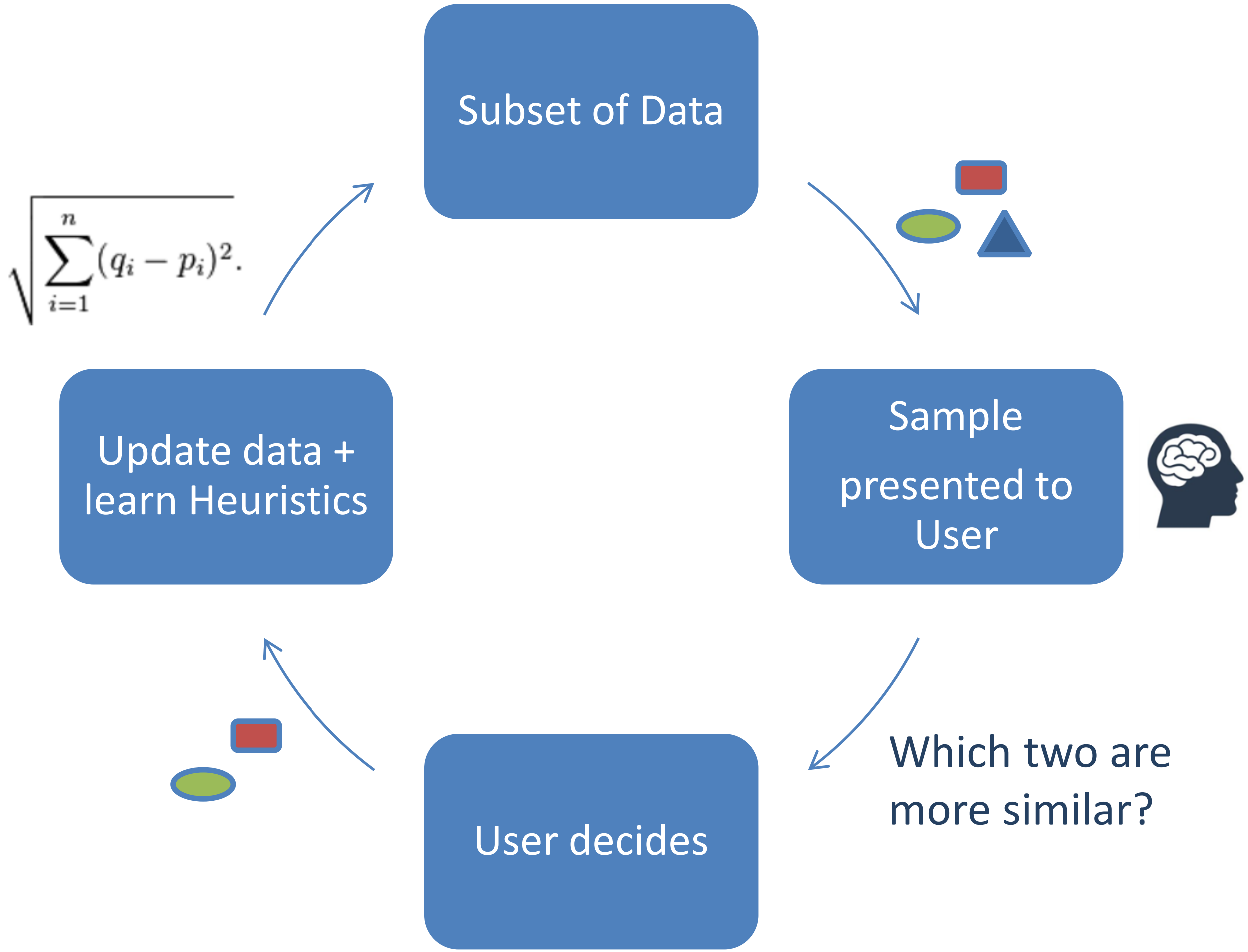
- Algorithm "A" => In real world it's a pipeline
- Task T => Prediction, Clustering, Classification, DimRed
- Performance P => TP, FP, Precision, Recall, F1, ...
- Experience E => Two general factors:
  1. More time
  2. More data => better data !!!

Source: Mitchell, T.M., 1997. Machine learning.

interactive Machine Learning Lifecycle

Case: data similarity:





Subset of Data

Sample presented to User

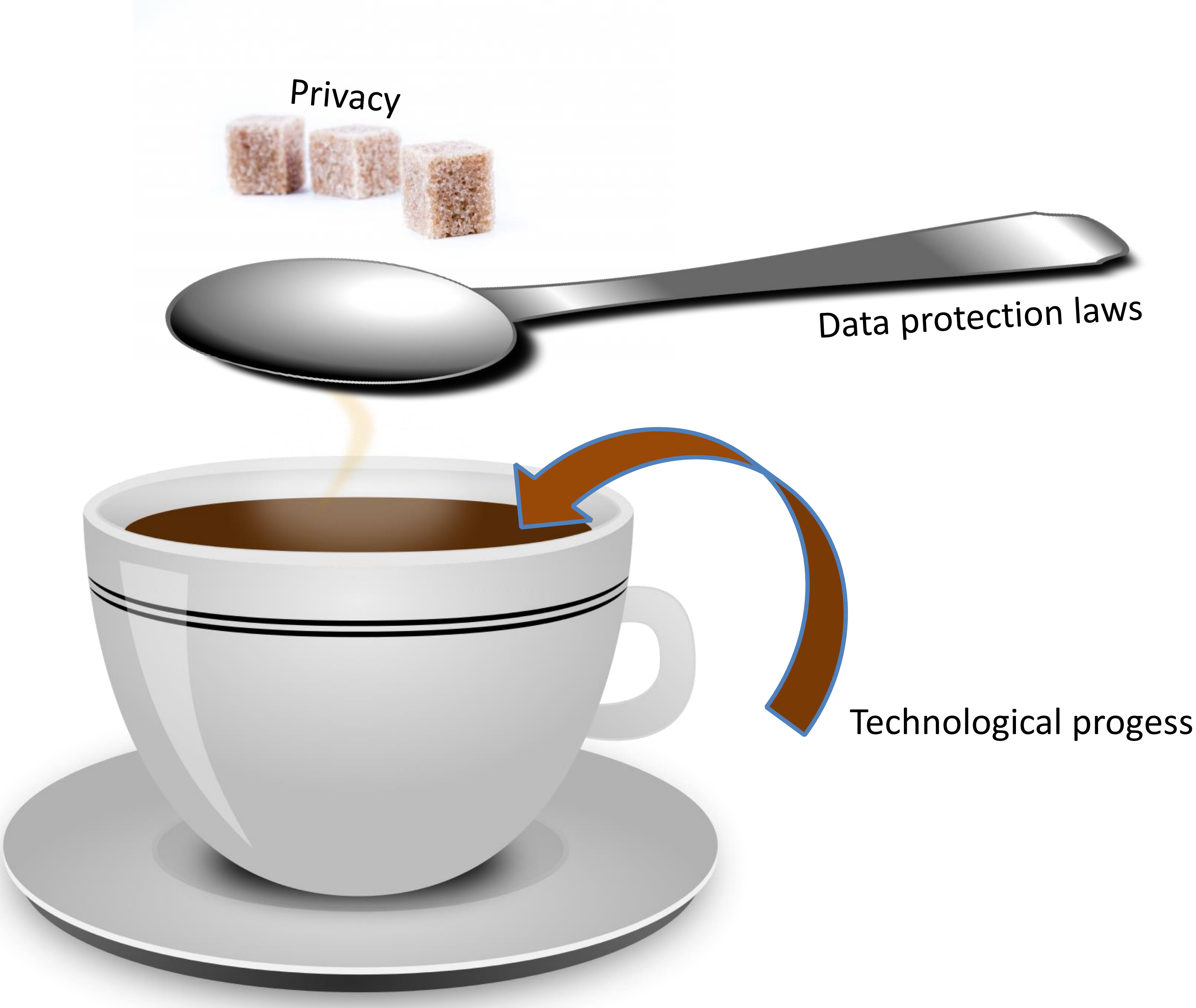
Which two are more similar?

User decides

Update data + learn Heuristics

$$\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Privacy in the 21st century… ??



Properties & General Approach

Data properties => Reduce granularity

Name	Age	Zip	Gender	Disease
Alex	25	41076	Male	Allergies
…	…	…	…	…

- Identifiers := immediately reveal identity
  - name, email, phone nr., SSN

  => DELETE

- Sensitive data
  - medical diagnosis, symptoms, drug intake, income

  => NECESSARY, KEEP

- Quasi-Identifiers := used in combination to retrieve identity
  - Age, zip, gender, race, profession, education

  => MAYBE USEFUL
  => MANIPULATE / GENERALIZE

K Anonymization criterion

k-anonymity: for every entry in the DS, there must be at least k-1 identical entries (w.r.t. QI's) => this is 3-anon:

Node	Name	Age	Zip	Gender	Disease
X1	Alex	25	41076	Male	Allergies
X2	Bob	25	41075	Male	Allergies
X3	Charlie	27	41076	Male	Allergies
X4	Dave	32	41099	Male	Diabetes
X5	Eva	27	41074	Female	Flu
X6	Dana	36	41099	Female	Gastritis
X7	George	30	41099	Male	Brain Tumor
X8	Lucas	28	41099	Male	Lung Cancer
X9	Laura	33	41075	Female	Alzheimer

Node	Age	Zip	Gender	Disease
X1	25-27	4107*	Male	Allergies
X2	25-27	4107*	Male	Allergies
X3	25-27	4107*	Male	Allergies
X4	30-36	41099	*	Diabetes
X5	27-33	410**	*	Flu
X6	30-36	41099	*	Gastritis
X7	30-36	41099	*	Brain Tumor
X8	27-33	410**	*	Lung Cancer
X9	27-33	410**	*	Alzheimer

Limits of anonymization

Trade-off between:

- Data utility => min. information loss
- Privacy => max. information loss

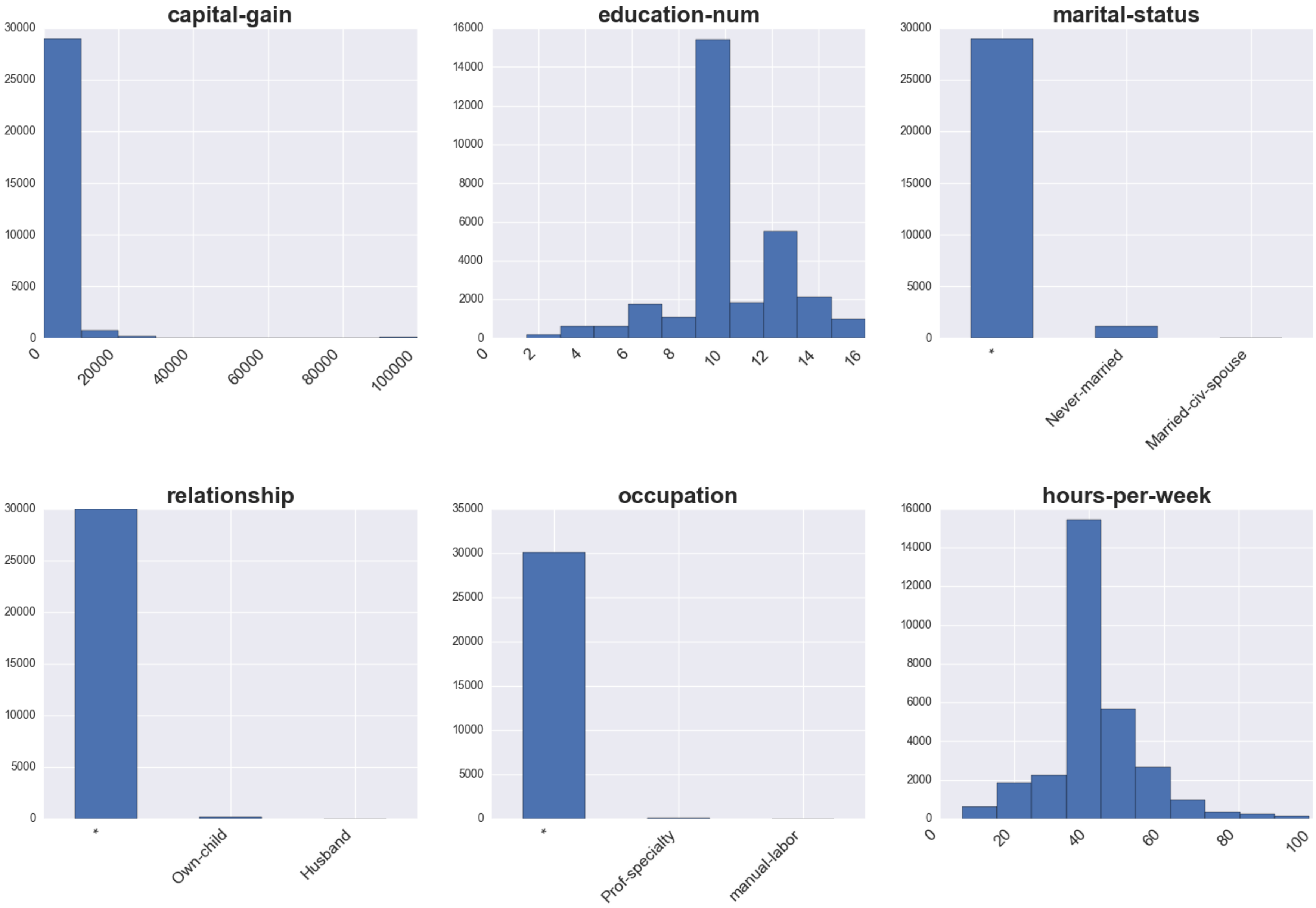
Both can be easily achieved (but not together ☺)

Node	Name	Age	Zip	Gender	Disease
X1	Alex	25	41076	Male	Allergies
X2	Bob	25	41075	Male	Allergies
X3	Charlie	27	41076	Male	Allergies
X4	Dave	32	41099	Male	Diabetes
X5	Eva	27	41074	Female	Flu
X6	Dana	36	41099	Female	Gastritis
X7	George	30	41099	Male	Brain Tumor
X8	Lucas	28	41099	Male	Lung Cancer
X9	Laura	33	41075	Female	Alzheimer

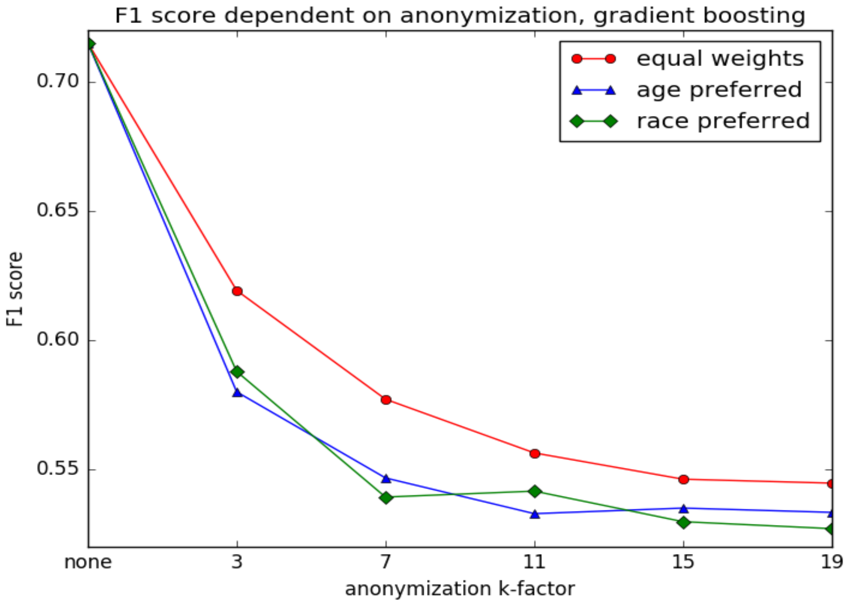
⇔

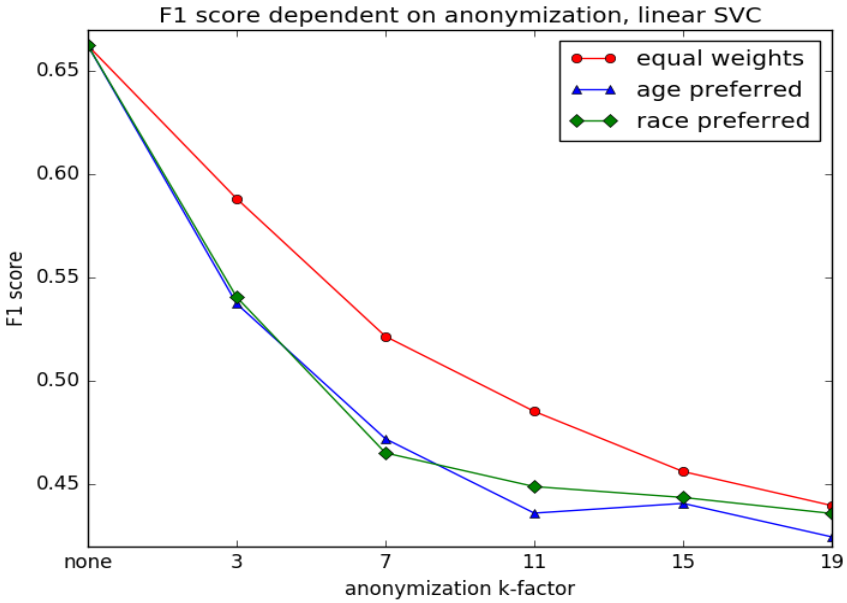
Node	Age	Zip	Gender	Disease
X1	*	*	*	Allergies
X2	*	*	*	Allergies
X3	*	*	*	Allergies
X4	*	*	*	Diabetes
X5	*	*	*	Flu
X6	*	*	*	Gastritis
X7	*	*	*	Brain Tumor
X8	*	*	*	Lung Cancer
X9	*	*	*	Alzheimer

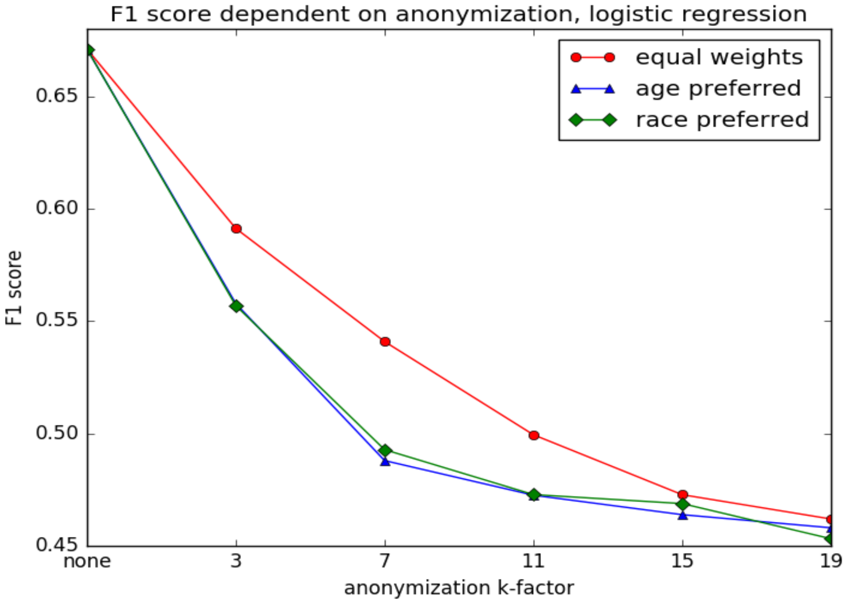
Anonymization - datasets

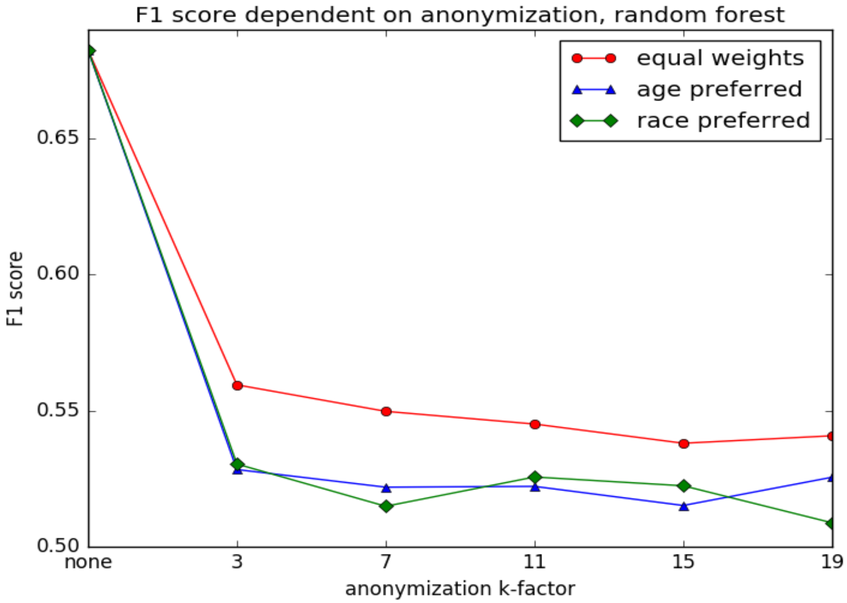


ML on Anonymization - Results









Adapting Weight Vectors 1/2

[51 - 76]	*	North_America	Male	*	Married-civ-spouse
[51 - 76]	*	North_America	Male	*	Married-civ-spouse
[51 - 76]	*	North_America	Male	*	Married-civ-spouse

57 | Private | United-States | Male | White | Married-civ-spouse

[48 - 70]	Private	America	Male	White	*
[48 - 70]	Private	America	Male	White	*
[48 - 70]	Private	America	Male	White	*

Adapting Weight Vectors 2/2

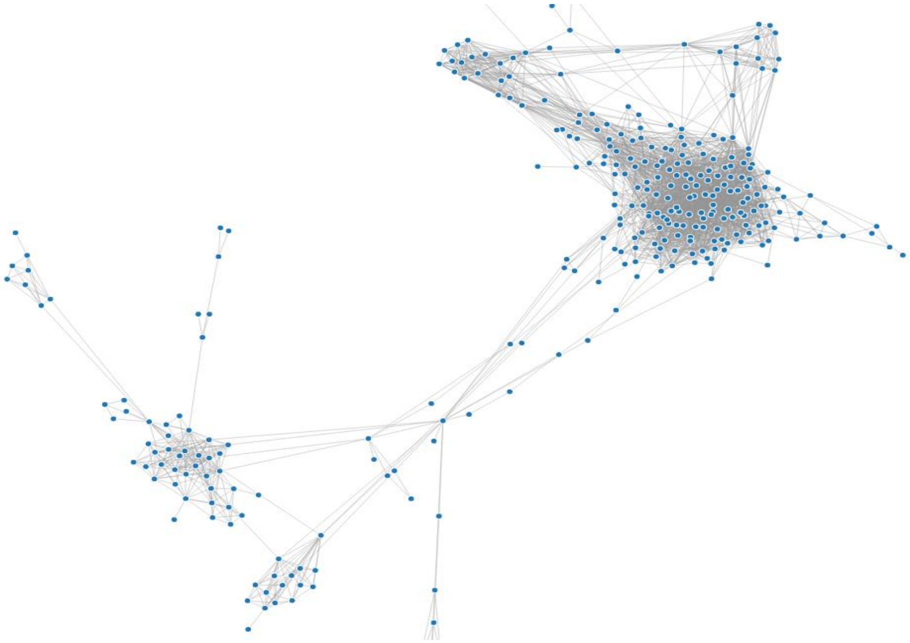
Applying a weight vector to our desired columns will change our cost function and thereby produce different anonymization results:

age	workclass	native-country	sex	race	marital-status
0.1667	0.1667	0.1667	0.1667	0.1667	0.1667

age	workclass	native-country	sex	race	marital-status
0.95	0.01	0.01	0.01	0.01	0.01

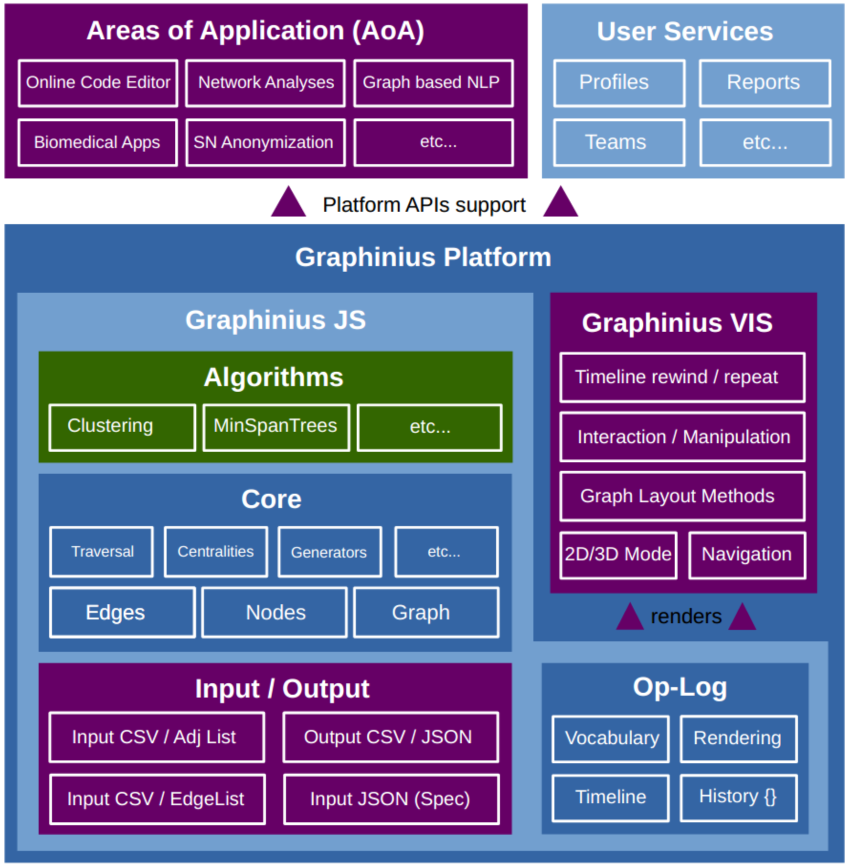
Graph based data sets

- Graph data / social network data, in which
  - nodes represent microdata
  - edges represent their structural context
  - graph data are harder to anonymize
    - It's harder to model the background knowledge of an attacker.
    - It is harder to quantify the information loss of modifications.



What is Graphinius?

- Graphinius JS => Graph library in Typescript (=> JS)
- Graphinius VIS => WebGL-based library



Areas of Application (AoA)

Online Code Editor | Network Analyses | Graph based NLP

Biomedical Apps | SN Anonymization | etc...

User Services

Profiles | Reports

Teams | etc...

Platform APIs support

Graphinius Platform

Graphinius JS

Algorithms

Clustering | MinSpanTrees | etc...

Core

Traversal | Centralities | Generators | etc...

Edges | Nodes | Graph

Graphinius VIS

Timeline rewind / repeat

Interaction / Manipulation

Graph Layout Methods

2D/3D Mode | Navigation

renders

Input / Output

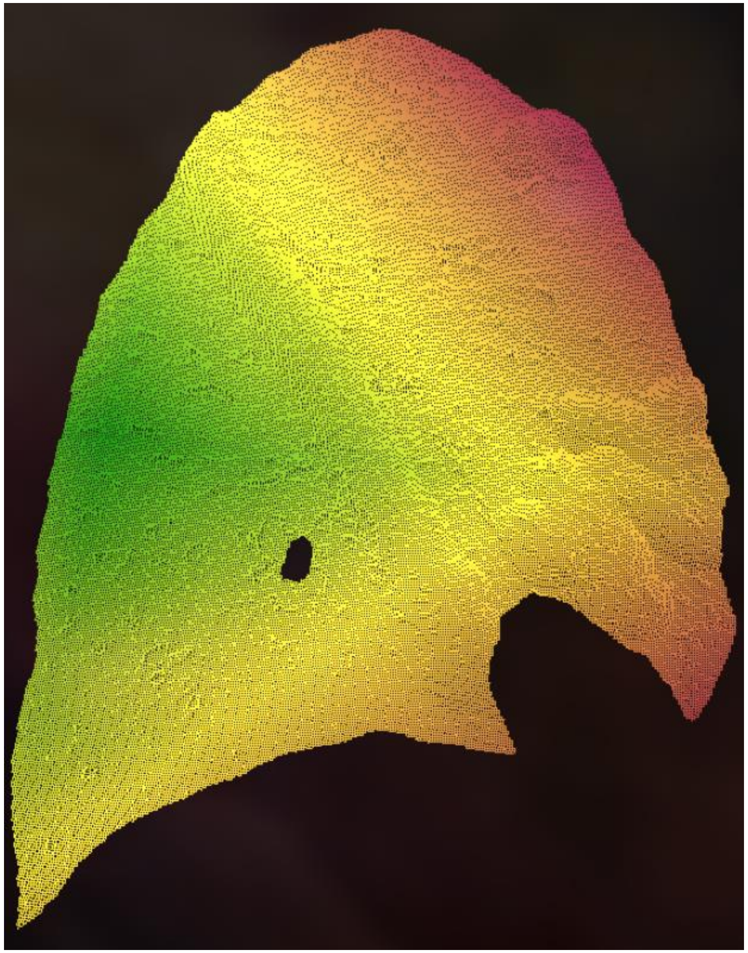
Input CSV / Adj List | Output CSV / JSON

Input CSV / EdgeList | Input JSON (Spec)

Op-Log

Vocabulary | Rendering

Timeline | History {}



Structure of experiments in AK-HCI

1. Write a simple UI in React / Angular (2)
2. Include the Graphinius JS library
3. Include the Anonymization JS library
4. Perform tests according to slide 11 ;)
5. We then compare the results..
6. If interesting / hard enough => write a report
7. Else => extend to graph-based structures / social networks

Thank you!